

## LA-UR-21-21059

Approved for public release; distribution is unlimited.

Title: The archives of the future

Author(s): Ali, Alee Rizwan  
Spivey, Whitney Jackson

Intended for: Web

Issued: 2021-02-04

---

**Disclaimer:**

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

# The archives of the future

BY RIZWAN ALI | FEBRUARY 03, 2021

NATIONAL ★ SECURITY SCIENCE

[ALL NEWS](#) / [PUBLICATIONS](#) / [NATIONAL SECURITY SCIENCE](#) / [WINTER 2020](#) / THE ARCHIVES OF THE FUTURE

Whitney Spivey | Editor

Artificial intelligence is the solution to digitizing, cataloging, and searching nearly 80 years' worth of classified materials at Los Alamos National Laboratory.

The term “[artificial intelligence](#)” (AI)—essentially programming machines to think like humans—conjures up different emotions in different people. Some view it with fear, imagining a malevolent AI similar to [Skynet in the Terminator movies](#). Others see something benevolent, such as the character [Data in Star Trek: The Next Generation](#). Still others see it as a means to advance the frontiers of science, engineering, and technology to new levels not possible through traditional means. At the National Security Research Center (NSRC or the Center), which is the classified library at Los Alamos National Laboratory, we see AI as a tool to help us go through the monumental tasks we have in digitizing, cataloging, and searching our collections.

Broadly speaking, AI involves developing smart machines that demonstrate human intelligence and cognition. The current state of AI is nowhere near anything we can classify as having sentience similar to Skynet or Data, and it's anyone's guess if machine consciousness will ever happen. Speculation about that possibility might be best left to futurists and science fiction authors.





Smart devices, such as speakers and phones, use AI to "learn" about their users. Dreamstime

However, a branch of AI, called machine learning, has made significant progress over the past couple decades. Machine learning, sometimes written as AI/ML, uses algorithms to recognize relationships in data. The algorithms, or sets of instructions to perform a certain task, “learn” from data rather than using a predetermined equation. An example is image recognition on smart devices that can identify who we are.

In fact, many of us use AI/ML on a daily basis and don’t think twice about the technological sophistication needed for devices and services to learn our behavior and make accurate predictions about our preferences. Digital assistants, such as Siri and Alexa on our smart devices, use AI/ML to learn the types of news we like to listen to and locations we typically travel to on the weekends, among other preferences in our routines. YouTube uses AI/ML to deliver the most-relevant videos for you based on your recent watch and search history. Smart thermostats know when we usually come home from work, adjusting the temperature so we arrive at a warm, toasty house in the cold months.

One significant limitation with AI/ML systems, however, is that they are very specific in terms of what function they can perform. For example, there is no chance, at least for now, that a smart thermostat’s AI/ML algorithm could be used to predict stock market trends. An AI/ML system must be developed and taught the specific set of tasks for which it was designed.

Those tasks, however specific, are quite remarkable and were not possible just a generation ago. This is why the NSRC is exploring this technology—to revolutionize the way we operate, and more importantly, the way we contribute to our nation’s security.





NSRC collections by subject. Los Alamos National Laboratory

## Machine learning and the Los Alamos mission

The NSRC [opened its doors in 2019](#), transitioning from what was a repository of archival materials to a dynamic, vibrant library that researchers regularly access. Already, the nascent NSRC is one of the largest research libraries in the United States.

The NSRC's specialized team of 40 historians, archivists, librarians, and digitizers partner with the Lab's scientists and engineers as they conduct research. They also curate the reports, films, photographs, lab notebooks, engineering drawings, and more that led to the dawn of the [Atomic Age](#). These early collections are among the NSRC's tens of millions of materials, which span the entire history—[more than 75 years](#)—of the nuclear enterprise. These historical documents and artifacts, classified research, and weapons information do not exist anywhere else.

Implementation of AI/ML technology in the NSRC is a priority for its leadership team so that we can better serve our researchers—the scientists and engineers at Los Alamos whose [work supports national deterrence](#). Nuclear weapons physicists, engineers, and production specialists use the collections daily in support of weapons' design and development to help maintain our nation's reliable and effective nuclear weapons stockpile. The materials are far from antiquated and dusty.

However, fewer than 10 percent of the holdings have been digitized, and fewer than 10 percent of those digitized holdings have been cataloged. This affects the speed with which the Center is able to provide researchers with the Lab's one-of-a-kind materials that are so vital to their work. Without employing AI/ML technologies in multiple areas of the Center, it is unlikely the NSRC will make significant progress in digitizing and cataloging our collections.

Some may be asking why it's important for us to digitize, catalog, and make searchable this vast collection of nuclear weapons material. The short answer: It saves a lot of time and even more money. Implementing AI/ML is a mission-critical task for the Center's collections to be accessible to researchers. The collections have reports and analyses that save the Lab tens of millions of

dollars annually because they preclude countless hours in redundant research, studies, and experiments. Furthermore, the majority of these records do not exist elsewhere; if researchers need them, the Center is the only option to get them.



NSRC collections by media type. Los Alamos National Laboratory

## Modernizing equipment and processes

The NSRC is exploring a variety of AI/ML technologies to digitize our vast collections of physical material, to automate tasks to capture metadata and catalog the digitized information, and to implement a natural (colloquial)-language search system. This will make digitized documents easier for researchers to find.

This technology is not entirely new to us. In 2020, we piloted an AI/ML system to digitize some of the documents in our microfilm and microfiche collections. These collections contain information relevant to nuclear weapons modeling and simulation, weapons designs, and plutonium pit production, which are a key part of warheads and must be replaced as they age. This work is critical to the Lab's stockpile stewardship mission (ensuring a safe, effective nuclear deterrent in the absence of weapons testing) and pit production benchmarks. Now, we want to extend our application of AI/ML technologies even further.

The Center's microfiche and microfilm number in the hundreds of thousands and contain well over 50 million pages of information. Using our current, non-AI/ML-capable equipment, software, and processes, it would take us an estimated 90-some years to digitize the microfiche collections, and more than 2,000 years to digitize our microfilm collection. The absence of AI/ML means significant labor is required to operate the outdated equipment and perform the cumbersome, manual quality-assurance (QA) process required for each digitized page. In the manual QA process, once the microfiche or microfilm is digitized, every single page needs to be reviewed to ensure the focus, contrast, alignment, proper resolution, and other factors were adjusted properly to ensure each page was readable. Each microfiche sheet can have nearly 100 pages and each microfilm reel can have up to 4,000 pages. The process to do this manually and adjust each page

on a single microfilm reel, for example, could take several weeks. With the AI/ML-based system, the process takes less than a minute. The computer automatically performs nearly all necessary adjustments and only flags a small handful of pages that the operator needs to review.

Modern AI/ML-based equipment and software, coupled with improved processes, has a high likelihood of reducing the amount of time necessary to digitize this material to less than two decades. The AI/ML systems could decrease the time to digitize the microfiche collection by as much as 80 percent and the microfilm collection by as much 99 percent by automatically detecting individual frames and performing highly sophisticated image corrections to automatically flag the dozen or so images out of several thousand that the AI/ML system was not able to automatically correct. This dramatically reduces the time our six archivists spend performing QA reviews on the finished digitized products.

But digitizing and performing QA on the documents are just two of several steps where AI/ML can be employed. The ultimate goal is not to just digitize the material, but to present researchers with materials that have been cataloged and are easily searchable using a natural-language search system.



John Moore of the National Security Research Center. Los Alamos  
National Laboratory

## A solution to a 400-year backlog

Once the content is available digitally, it can be cataloged. Currently, cataloging in the Center is a manual, time-consuming process where our librarians or archivists upload metadata information into one of our classified digital repositories. The metadata contains information such as the document's title, date, author(s), report number, organization, abstract, and keywords. This process can take anywhere from 10 to 30 minutes per document depending on the complexity of the document and speed of the system that particular day. However, if the information isn't cataloged, it isn't discoverable to researchers.

If the NSRC continued to manually catalog its current backlog of 2.4 million digitized documents, it would take more than 400 years to complete. Meanwhile, as we begin to increase the rate at which we are digitizing our physical collections, the total number of digitized documents will continue to grow at a rapid pace. The NSRC doesn't have enough staff to manually catalog these vast collections of digitized documents.



To automate the cataloging process, an AI/ML-based system needs to be implemented. This system would perform a sophisticated optical character recognition process and parse the information in each document into metadata, which could then be cataloged. After the metadata is parsed, it would pass the information to an AI/ML-based, natural-language search system. These distinct processes will involve the NSRC's partnership with several companies that specialize in AI/ML.



AI will revolutionize the way the NSRC operates. Los Alamos National Laboratory

The NSRC's documents date back more than 75 years to the inception of the Manhattan Project and contain older typewriter fonts that cannot be searched using industry-standard document viewing software, such as Adobe Reader. The AI/ML system needs to read each document regardless of its format and fonts and then extract the required metadata so it can be cataloged. The metadata extraction system scans the digitized documents and, through an AI/ML process, teaches itself where to find the relevant metadata information.

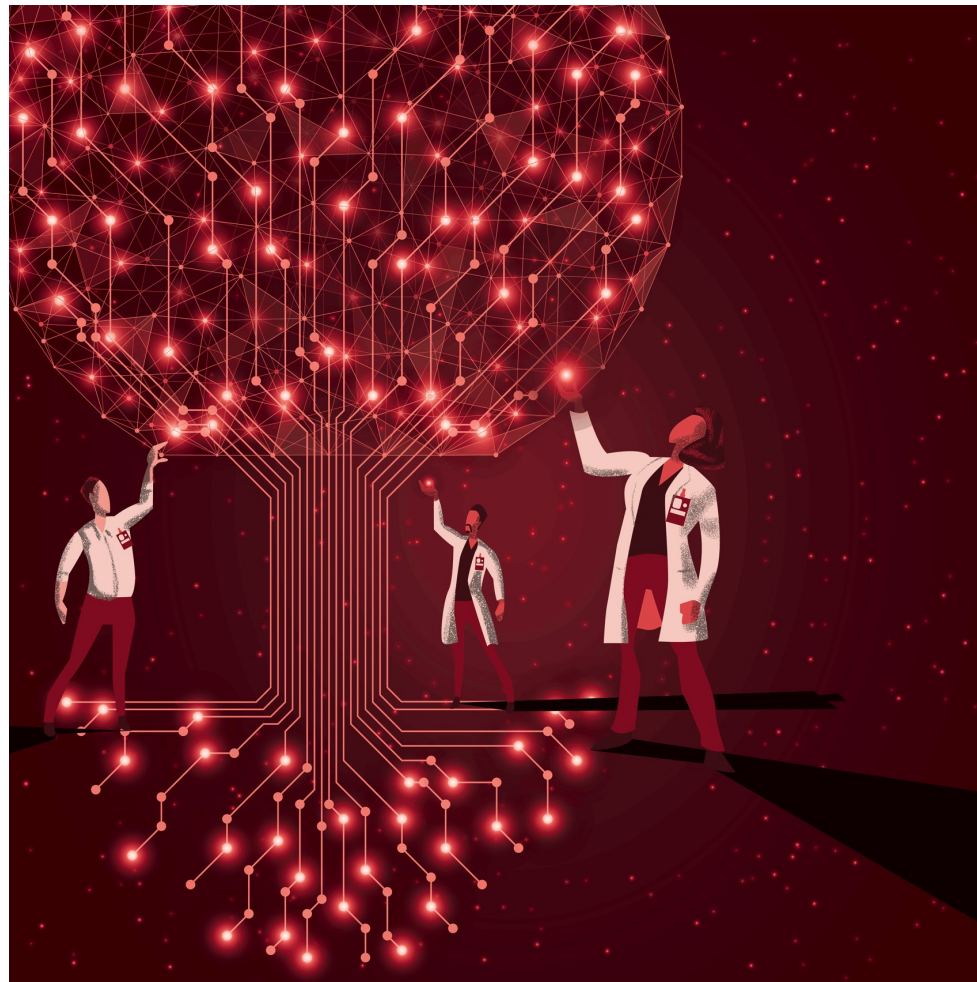
This information is then passed to another system that uses AI/ML to implement a natural-language search tool. The natural-language search tool uses contextual clues in the way a researcher would phrase a search query to discern the intent of the search, rather than deliver just a simple list of documents that contain the words in the search field.

To illustrate this point, suppose a researcher wanted information about the word "plant." The system would use contextual clues in the full search string to determine if the researcher wanted to know about the biological entity "plant," a manufacturing "plant," or how to "plant" something in the ground. Each meaning would yield vastly different results, and the use of an AI/ML-based natural-language search system would deliver only the most relevant results to the researcher.

To address this metadata/search challenge, the NSRC initiated a large-scale AI/ML project to automatically catalog our digitized information and provide a natural-language search system to help researchers find relevant information. Companies that do this highly specialized type of

AI/ML are rare. To find these companies, the NSRC reached into the U.S. Intelligence Community, which has a very similar problem set—namely that it has vast quantities of digital information to catalog and search in a rapid and efficient manner.

After a fairly lengthy process, the NSRC found a set of companies that specialize in using AI/ML to extract metadata from digitized documents and use natural-language, AI/ML-based systems to search through materials.



AI will be used to digitize and catalog information; once information is catalogued, it is easily searchable. Los Alamos National Laboratory

## A successful test run

[Bob Webster](#), the deputy Laboratory director for Weapons, provided the NSRC funding to test the system in a six-month pilot study on the Lab's unclassified network, using unclassified digitized nuclear power plant material. The system's AI/ML system successfully captured the required metadata automatically, to include keywords, and populate the cataloging system. Because this system was also used within the Intelligence Community, it passed the Lab's classified analysis, which confirmed it can be installed and used securely on the Lab's classified network.

The end goal for this AI/ML installation on the Center's classified network, an initiative called Titan on the Red, is to extract metadata from various digital data repositories and present researchers with a natural language, AI/ML-based interface to search through the NSRC's entire digitized collection. Because many of the documents within the NSRC are protected through stringent security and need-to-know protocols, the search system will enforce these protocols and only deliver documents that the researcher has the appropriate approvals to view.


What this means for researchers is that the large backlog of documents that are currently not cataloged and not searchable will become accessible in a matter of months once the system comes online, rather than in double-digit decades. Plus, the process to search through the NSRC's digital repositories will become dramatically easier than the current process, which requires contacting one of our librarians to have him or her manually search through our collections.

For the 2021 fiscal year, the Center’s goals are to install Titan on the Red on the Lab’s classified network and begin integrating the system into at least one of the digital repositories. Additionally, the Center intends to begin the process of training the AI/ML to extract necessary metadata from the documents as well as to identify words and terms specific to our collections.

In the beginning, the system will only be available to the NSRC’s staff and a select group of researchers, as we fully test the system. The eventual goal is to make the system available to everyone in the Lab’s Weapons program and others who have a need to access Weapons program material.

AI/ML is new, yet proven. The Laboratory needs to embrace this advancement, which is really the only solution to making its one-of-a-kind collections searchable to its researchers. Investing in AI/ML saves countless hours and many millions of dollars, while directly contributing to the Lab’s mission success and our nation’s security.

Powering Tomorrow's Innovations Today: National Security Science



The NSRC powers tomorrow's innovations today. Los Alamos National Laboratory

# About Rizwan Ali





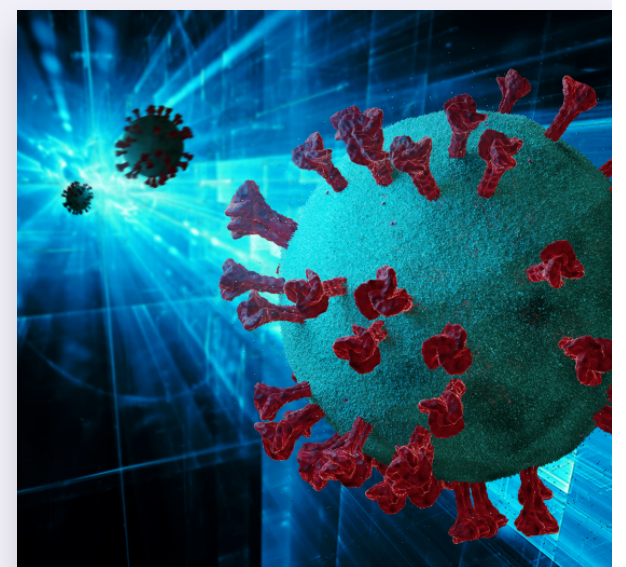
Rizwan Ali is the director of the National Security Research Center (NSRC), the classified library at Los Alamos National Laboratory. As the NSRC director, he sets the strategic direction for one of the largest libraries in the United States.

Ali brings with him 30-plus years of military, national security, cyber security, nuclear weapons, and engineering experience. He is a retired U.S. Air Force colonel, whose career included commanding multiple, large units in the United States and in combat zones. He has extensive international work experience with NATO and U.S. partners in the Middle East and Africa.

Ali has a bachelor's degree in engineering from Stevens Institute of Technology; a master's degree in public administration from Troy University; a master's degree in military history from Air University; and a master's degree in international relations from National Defense University.

## MORE STORIES

[NSS Home](#) >



Entering the realm of augmented reality

[Read More](#) >

Raiders of the lost archive

[Read More](#) >

Computing for a cure

[Read More](#) >



**Los Alamos**  
NATIONAL LABORATORY

Los Alamos National Laboratory  
P.O. Box 1663  
Los Alamos, NM 87545  
(505) 667-5061



AT THE LAB

- [Business Opportunities](#)
- [Jobs](#)
- [Organizations](#)
- [Research Library](#)
- [User Facilities](#)

INFORMATION

- [Emergency, Fire](#)
- [Events, Lectures](#)
- [Ombuds](#)
- [Resources](#)
- [Reading Room](#)
- [Science Museum](#)

FOR EMPLOYEES

- [AskIT](#)
- [LANLINSIDE](#)
- [MyMail](#)
- [New Hire Process](#)
- [SSL Portal](#)
- [Training](#)



[Contact Us](#) | [Terms of Use/Privacy](#) | [Site Feedback](#)

Managed by [Triad National Security, LLC](#) for the [U.S. Dept. of Energy's NNSA](#) ©Copyright [Triad National Security, LLC](#). All Rights Reserved.